Universe
Scientific Publishing

# The research and implementation of DNA sequence minimization tree based on Spark cloud computing and chaotic genetic algorithm

**Qingxue Liu**

**Jilin University of Architecture and Technology, Changchun 130012, China.**

***Abstract:*** to reconstruct a reliable phylogenetic inference based on the genetic or species diversity of living organisms, and to reveal the sequence of biological evolution, is helpful to understand the history and evolutionary mechanism of biological evolution. In order to solve the problem of limited running time and number of categories in the existing algorithms for constructing maximal reduced tree, chaos genetic algorithm and Spark cloud computing are used to parallel the algorithm, which can greatly reduce the searching time, increase the number of treated species.

***Keywords:*** Maximum Minimalistic Tree; Chaos Genetics; Spark Cloud Computing

## Introduction

The bioinformatics is a new cross-discipline combining life science and computer science. It focuses on the collection, computation, storage, analysis and interpretation of biological information data. To reconstruct the evolutionary history of organisms and express the evolutionary relationship among groups in the form of phylogenetic tree is always the core problem of phylogeny and one of the important contents of evolutionary biology. The process of constructing phylogenetic tree is the process of inferring the evolutionary history from the information of biological sequence and“Reconstructing” the phylogenetic relationship, the evolutionary relationship is expressed in the form of evolutionary tree-the leaf node of the tree represents each biological sequence, and the length of the branch represents the evolutionary distance between organisms. Iterative methods commonly used in search algorithms can be implemented using genetic algorithms, which are easy to parallelize to meet current trends in low-cost clusters and multi-core processor, however, there are some problems such as local optimization and slow convergence, and the traditional parallel method can not deal with the data effectively, while chaotic genetic algorithm can realize population diversity and rapid convergence, spark cloud computing can greatly save read and write time, greatly improve the efficiency of computing.

## 1. Basic algorithms

The Maximum parsimony is to perform this calculation on all possible correct topologies and select the topology with the smallest number of substitutions as the optimal occurrence tree. The common heuristic to find the maximal minimalistic tree is the stepwise addition method. The initial tree of 3 species was built first, and then the fourth species was inserted into 3 branches of the initial tree. The length of the tree was calculated by MP method, and the minimum of the 3 tree lengths was recorded. This process is repeated until a tree containing all species is created, producing a temporary MP tree. Then we use the branch-exchange strategy to find a tree with smaller length, and then stop the exchange under certain conditions to get the maximum reduced tree.

Genetic algorithm (GA) is an adaptive probability search algorithm for global optimization, but it has some problems, such as premature convergence and random roaming, which bring inconvenience to its application. The basic idea of chaos is to transform the unknown variable from the original chaotic space to the solution according to certain rules. The chaos optimization method is sensitive to the initial conditions, it is easy to achieve global asymptotic convergence.

The combination of chaos optimization theory and genetic algorithm can make up for the disadvantage of slow convergence speed and easy falling into local optimum of genetic algorithm. In the process of chaos optimization, the initial individuals are first generated by using chaos theory, and then a small chaotic disturbance is added to the center of the best or worst point obtained by genetic algorithm to search for

them in more detail, thus out of the local optimization, to find the global optimal point.

## 2. Research on the algorithm of constructing maximum reduced tree based on chaos genetic algorithm

By analyzing the traditional genetic algorithm, the chaotic genetic algorithm based on Logistic map is proposed, and the chaotic sequence generated by chaotic map is used to replace the process of random individual generation to improve the genetic algorithm, it can keep the diversity of the population, improve the efficiency of the algorithm and reduce the number of iterations to reach the global optimal solution. The following issues need to be addressed:

1) the genetic code of the evolutionary tree, since the input data is a nucleotide sequence, composed of a, C, G, T (U) , so directly use these four letters, the input of each nucleotide sequence as a code, no additional action is required.

2) Design of fitness calculation function. In order to speed up the convergence of the algorithm, we define the historical maximum reduced tree as the minimum adaptive tree which appears in the whole searching process, value.

3) Design of chaotic genetic operators. According to the practical problems to be solved, three kinds of genetic operators, selection, chaos crossover and chaos mutation, are adopted in the design of chaos genetic operators.

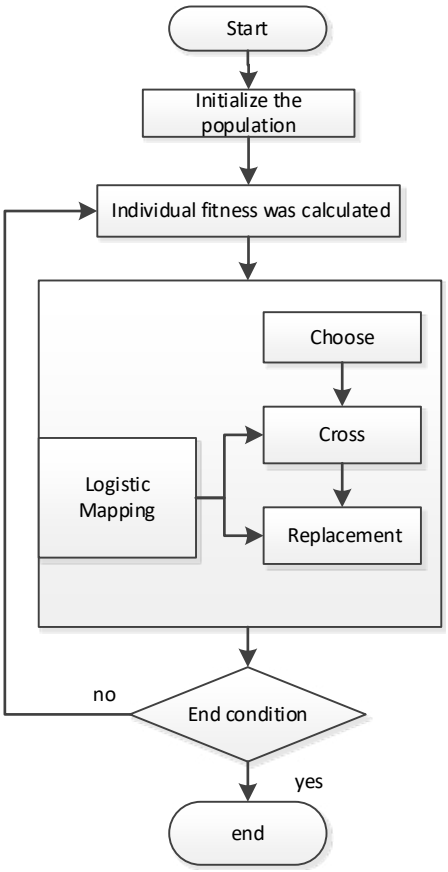


Fig. 2 flow of chaos genetic algorithm

## 3. Spark cloud computing based on the largest minimalistic tree construction of the implementation

Based on the Spark cloud computing, the implementation of the maximal minimization tree is realized by using TreeBASE data source, using the virtualized Spark cloud computing platform to manage the configuration and monitor the running services of the entire virtualized cloud computing platform through the physical module, virtual cloud computing platform can support nearly 1,000 virtual machine allocation applications, and realize the parallel computing of maximum reduced tree construction.

### 3.1 Technical difficulties and key technologies to be solved

1. Aiming at the problem of premature and slow convergence speed of traditional genetic algorithm, the chaos theory is introduced into the design of genetic operator, and the chaotic sequence generated by chaotic map is used to replace the process of random individual generation to improve genetic algorithm, it improves the efficiency of the algorithm, keeps the diversity of the population and reduces the number of iterations to search the global optimal solution.

2. In order to solve the problems of limited memory space and low running speed in traditional parallel computing, hDFS distributed data store are proposed for data storage and analysis to increase the reliability of data acquisition and the speed of data query and retrieval, and Spark cloud computing to increase speed.

### 3.2Forecast and analysis of main technology application and formation

This project is to carry out the work by adopting the overall technical route of the theoretical research combined with the cross-platform construction, the specific technical route. Based on theoretical research, the advantages and disadvantages of related algorithms at home and abroad are deeply analyzed, and various methods are combined to complement each other. First of all, the study of Maximum parsimony, clear the construction of a simple tree, tree length calculation. The genetic algorithm is studied, and the coding mode, the calculation function of tree length and the establishment of genetic operator are made clear. The chaos optimization algorithm is studied, and the combination mode of chaos genetic algorithm is defined. Research Big Data Theory and platform building technology, Clear Hadoop cluster deployment, Spark cluster deployment and development environment installation, ready for the experiment. Extract sequence data from the TreeBASE library as a source of test data. To prepare for later experiments to extract genes. And the results of the theoretical research process published papers。

## 4. Conclusion

Based on the molecular characteristics of species, phylogenetic trees were constructed to understand the relationship between species. The project builds phylogenetic tree based on Spark cloud computing and chaotic genetic algorithm, which enhances the sensitivity of searching and optimizes the local searching ability of genetic algorithm. Spark-based parallelization method can greatly improve the efficiency of comparison operation, using Hadoop HDFS as the genome data storage file system, the problem of scalable incremental storage of massive high-throughput genomic data is solved, and the number of treatable species is increased. The function and structure of macromolecules were analyzed. Macromolecules of the same family had similar tertiary structure and biochemical function. The phylogenetic tree was constructed by sequence homology analysis.

## References

[1] M. Fischer, S. Kelk, On the Maximum Parsimony distance between phylogenetic trees, Ann. Comb. 2016;20 (1) 87–113.

[2] Diep Thi Hoang ,A new phylogenetic tree sampling method for maximum parsimony bootstrapping and proof-of concept implementation. 2016 Eighth International Conference on Knowledge and Systems Engineering .

[3] Liu Zhenyu. Parallel Research of genomics data alignment algorithm based on Spark, and construction of alignment platform [ D ] . Hohhot: Inner Mongolia Agricultural University, 2019.

[4] Yu ru,An improved cloud adaptive genetic algorithm combined with chaotic search, Changchun Normal University. 2023,42(02)

Author Introduction: Liu Qingxue (1977-), female, Han nationality, master's degree, professor, engaged in research on intelligent algorithms and computer science and technology. Email: 406428957 @ qq.com

Supporting project number: Xiaokezi [2022] 009ZQKJ